Gene Hunting in Complex Traits

Lips, E.S.

2017

document version
Publisher's PDF, also known as Version of record

6

Summary & Future Directions

6.1 SUMMARY

6.1.1 *Summary of my PhD*

About ten years ago the hunt for genes that contribute to complex traits opened up, when it became technically feasible to assess hundreds of thousands common SNPs by means of a Genome-Wide Association Study (GWAS). However, due to the large number of tests typically conducted in GWAS, a stringent correction for multiple testing needs to be applied, and therefore real but weak associations are likely to be missed. Complex traits arise from a complex interplay between multiple genes (and environmental factors). Therefore, adding a priori knowledge on the biological interplay between genes to a genetic study may increase our changes to identify molecular mechanisms that contribute to the trait.

My PhD project aimed to develop novel statistical methods that could aid in the identification of genes and molecular mechanisms that underlie complex disorders and traits. The main objective was the development of a novel powerful gene-based and gene-set association method for GWAS data, to test hypotheses on the concerted actions of multiple genetic variants and genes.

In an unpublished pilot study, described in Chapter 2, I developed the first version of a statistical pipeline in which the effects of genotyped SNPs located in a set of biologically related genes can be simultaneously tested for association in a self-contained manner. To evaluate the performance of this novel approach, I conducted a gene-set analysis on 23 sets of biologically related presynaptic genes in two separate case-control samples in which cases were affected with Bipolar Disorder (BD): one sample contained 2053 unrelated European-American (EA) and the other sample contained 1018 African- American (AA) individuals. To test the performance of the developed approach, I compared the gene-set association results from the described gene-set analysis with two other suitable gene-set association approaches: the Mann-Whitney U test and set-based test in PLINK. Although no significant gene-sets were detected in the EA sample in any of the statistical approaches, I could show that all approaches show a similar pattern in *P*-value distribution over the different gene-sets in the EA and the AA sample. Furthermore, our novel approach and the gene-set based test in PLINK showed the largest number and overlapping significant gene-sets in the AA sample. Results from this study suggest that in the AA sample presynaptic sets of genes that have a role in the serotonin pathway, neurotransmitter synthesizing/metabolizing and

cytoskeleton are involved in BD.

Chapter 3 describes a published study (*Lips et al.*, 2012), in which I extended the algorithm as described in Chapter 2 with a competitive test. Subsequently, I applied both the self-contained test and the competitive gene-set test to schizophrenia (SCZ) on a sample that consisted of 9638 unrelated individuals of European ancestry: 4673 were diagnosed with SCZ and 4965 individuals served as controls. I could show that a set of 1026 synaptic genes was significant associated with SCZ. Apart from testing this large gene-set I tested 17 smaller sets of synaptic genes, which were grouped on the basis of their shared cellular function. From these smaller gene-sets, three also showed association with SCZ. The genes in these gene-sets are involved in intracellular signal transduction, excitability and cell adhesion and trans-synaptic signaling.

Following the successful application of my statistical pipeline to BD and SCZ, I developed a highly generic and freely available version of my algorithm, which included both the self-contained and the competitive gene-set association test. This command-line tool, which we named Joint Association of Genetic Variants (JAG), gives the user the full freedom to specify his own sets of biologically related SNPs (e.g. located in single genes, gene-sets or sets of regulatory elements) and other parameters. I presented this statistical tool to the public in a publication described in Chapter 4 and on our website (http://ctg.cncr.nl/software/) (*Lips et al.*, 2015) This study includes a power analysis, which is conducted on Crohn's Disease (CD). By comparing the performance of JAG's self-contained test with other available self-contained gene-set analysis tools, I showed that JAG had more power to correctly identify validated gene-sets for CD than other available tools.

In the final stage of my PhD project I investigated whether essential genes, monogenic diseases and/or complex traits have quantifiable characteristics that discriminate them from each other and that can aid in prioritizing (complex) disease genes. In this study, described in Chapter 5, I showed that the number of variants (which are corrected for length of the sequenced exome) is a function of the lethality of the gene: the lowest number of variants could be observed in essential genes, which are believed to be the most crucial genes for survival of an organism, and increases in number of variants via monogenic disease genes and complex disease genes to non-associated genes (in this order). This observation is in line with selection pressure: essential genes are more constrained to have mutations than genes that are less critical for survival. A similar trend is observed

by an analysis conducted on several network properties: essential genes tend to have a more central and more crucial role in the protein-protein interaction network (PPI) and monogenic genes and complex disease genes in between, whereas non-associated genes have the least critical role in the PPI and tend to be located at the periphery of the PPI network. Results from this study suggest that gene properties could be a potential source to predict the role of a gene in disease.

6.1.2 *Summary of the GWAS field*

Before discussing the future directions of the identification of genes and molecular mechanisms that are involved in complex traits, I would like to address the most important lessons we learned after a decade of GWAS analyses.

Lesson 1: GWAS works

GWAS has proved to be a successful method to detect genetic variants that associate with complex traits. However, we have also learnt that a significant genetic variant in GWAS does not likely reflect a true causal variant. Since there are many more genetic variants in the population (∼11M) than the number of SNPs in a GWAS (∼0.5M), it is more likely that a significant SNP is a tagging SNP, which is in high LD with a variant that has a true (in)direct effect on the phenotype, rather than being the causal SNP (*Visscher et al.*, 2012). This means that a significant GWAS-hit is not an end-point, but rather a starting-point for in-depth in silico annotation and possible resequencing. To date, March 2016, hundreds of successful GWASs together have yielded more than 23,000 SNPs with P-value $<$ 10e-06, implicating more than 3300 genes for investigated traits and diseases. Although for many traits a polygenic nature was already assumed in 1910 (*East*, 1910), it has become explicitly evident in the GWAS era that many complex disorders/traits are highly polygenic: they are likely caused by genetic variants located in thousands of genes. Most of the identified variants in GWAS show at most a 1.1-1.5 increase in risk and thus have a small effect on the trait under investigation (*Bodmer & Bonilla*, 2008; *Lander et al.*, 2001). SNPs with even smaller effects are likely to be missed due to the stringent P-value threshold that is applied.

Lesson 2: The heritability is both hidden and missed by GWAS

A few years after GWAS was introduced, it became evident that there was a huge gap between the variance explained by the variants identified through GWAS and

the estimated genetic variance/heritability of the trait. The debate about whether this represented missing or just hidden heritability started. In a search for finding answers on where to look for this unexplained heritability several explanations were suggested (*Maher*, 2008).

One possible reason for the inability to explain the heritability was that so far GWASs were underpowered. The first wave of GWASs, which included only a couple of hundreds/thousands of individuals, showed that the sample sizes were underpowered to detect variants with smaller effect sizes than 1% of the variance. Variants with smaller effect sizes were expected to be identified when sample sizes would increase. Indeed, GWAS sample sizes have increased tremendously in the next years, from a couple of thousands to hundreds of thousand individuals, and identified many more variants with tiny effects. For example, a GWAS on height which included ∼63,000 individuals revealed 54 new genetic variants for height, which together explain only 5% of the heritability of height, which is estimated around 80% (*Visscher*, 2008). A subsequent GWAS on height, which included ∼184,000 individuals, reported 180 associated variants that account for 10% of the genetic variance (*Lango Allen et al.*, 2010). Similarly, a GWAS on Schizophrenia (SCZ) including ∼40,000 individuals found 5 associated genetic variants which were not previously reported and another 14 new genetic variants were identified when sample size increased to 60,000 individuals (*ISC*, 2011; *Ripke et al.*, 2013). A third meta-analysis on SCZ that included 150,000 individuals identified another 83 newly and independently associated variants, together explaining 3.4% of the estimated heritability of SCZ, which is estimated to be 80% (*Ripke et al.*, 2014). These studies have also revealed the polygenicity of these traits and that the genes involved are enriched in relevant molecular processes. These huge GWAS studies showed us that although larger studies detect more variants that explain a higher proportion of the trait-heritability, even a sample size $>$ 100,000 individuals does not close the gap between the variance that can be explained by associated variants and the estimated trait's heritability.

Since only a few common variants with small effect sizes could be identified for most complex traits, it was speculated that the unexplained heritability might be captured by rare variants (*Maher*, 2008). Although rare variation is occurring at a much higher rate in the human population than anticipated and complex traits are expected to arise from a combination of common and rare variants, it is also expected that most of the heritability is captured by common variants. As the

heritability of a trait is a function of the number of variants, the observed frequencies (MAF) and their effect sizes (*Wray et al.*, 2012), under a near neutral model and variants being distributed over a wide range of allele frequencies most of the heritability will be captured by common variants (*Visscher et al.*, 2012). It is due to the stringent significance threshold that common variants with a true but small effect on the trait are likely hidden in the GWAS data since they do not reach significance.

With the development of Genome-Wide Complex Trait Analysis (GCTA) it became possible to estimate the proportion of additive genetic variance that is tagged by all genotyped SNPs that are included in GWAS data (*Lee et al.*, 2011). When applied to GWAS data for height, including ∼300k genotyped SNPs and ∼4,000 individuals, about 45% of the heritability could be explained when considering the common genotyped SNPs. When applying GCTA on a SCZ sample, about 23-33% of the heritability could be explained by all genotyped SNPs (*Lee et al.*, 2011; *Ripke et al.*, 2013). Both examples show that the variance explained from all genotyped SNPs is roughly about the half of the trait's heritability as estimated in twin studies. This suggests that more causal common variants can be detected when using extremely large samples. In addition, a gap between heritability estimates from GCTA and the heritability estimates from twin studies leaves room for effects of rare variants (*Vinkhuyzen et al.*, 2013).

Lesson 3: Regulatory variants contribute to complex traits

In total, 47.5% of the SNPs identified through GWAS are located within the transcription start site (TSS) en transcription end site (TES) of a gene, which spans in total ∼35% of the human genome (*Maurano et al.*, 2012; *ENCODE*, 2012). Furthermore, the GWAS hits are enriched in coding regions: while about 3% of the human genome cover the protein-coding exons, ∼11% of the GWAS hits are located or in strong LD ($r^2 > 0.8$) with a SNP located within these coding regions (*Maurano et al.*, 2012). However, only 9% of the GWAS hits that are located within these coding regions are in strong LD with non-synonymous variants (*Hindorff et al.*, 2009). This suggests that the majority of variants underlying significant hits for complex disorders/traits do not lead to a change in the structure of the protein. This is in contrast to what is observed in monogenic disease genes, where the vast majority of the causal variants are non-synonymous/missense variants (with an OR > 2) and may imply that the genetic variants that associate with

complex traits have a role in the regulation of gene expression (*King & Wilson*, 1975; *Bodmer & Bonilla*, 2008).

In recent years, the non-coding variants that are influencing the expression levels of genes are identified by means of expression quantitative trait loci (eQTLs) mapping. In this type of study, GWAS data and gene expression data are assayed simultaneously and the expression pattern of individual genes between samples of individuals that are genetically different analyzed (see Albert et al. for a large overview of recent human eQTL studies (*Albert & Kruglyak*, 2015).
Furthermore, the Encyclopedia of DNA Elements (ENCODE) project released an update of the human genome annotation in which 80% of the human genome was annotated. This update included detailed information on mapped regions of transcription, transcription factor binding sites, chromatin structure and histone modification (*ENCODE*, 2012). By using this data, multiple studies have provided evidence for a regulatory role of common SNPs implicated in disease. For example, reported SNPs are significant enriched for cell type specific enhancers and deoxyribonuclease I hypersensitivity sites (DHSs) (*Ernst et al.*, 2011; *Schaub et al.*, 2012). Another study examined whether ∼5000 non-coding variants that previously have shown to associate with a trait (*Hindorff et al.*, 2009), are enriched in regulatory regions to which transcription factors bind (DHSs), and showed enrichment for SNPs located in DHSs as well as enrichment for SNPs that are in perfect LD with nearby DHSs (*Maurano et al.*, 2012). Results from a recent study showed that a substantial proportion of the heritability of complex disorders is captured by genotyped SNPs that are located in DHSs (*Gusev et al.*, 2014). Others showed that SNPs that associate with a given trait influence gene expression by altering chromatin marks within cell types that are relevant to the trait (*Trynka et al.*, 2013). This study and others also demonstrated that data on regulatory elements can aid the identification of the causal SNP by screening the SNPs that are in LD with the tagging SNP for overlap with regulatory elements (*Trynka et al.*, 2013; *Claussnitzer et al.*, 2015).

In conclusion, the first half of a decade of GWAS' can be seen as a period of initial successes, big debates and technical challenges, whereas the second half is one of maturing and great successes. Ten years of GWAS have shown that complex traits are (highly) polygenic in nature. And in recent years it became more evident that common variants substantially contribute to complex traits and diseases. About 45% of the GWAS hits point to genic regions, but in a different

fashion than observed in Mendelian diseases: the majority of the GWAS hits are located in intronic regions and only a few are in LD with non-synonymous variants. It also became evident that common (non-coding) regulatory variants have a substantial effect on the phenotype (*Maurano et al.*, 2012; *Gusev et al.*, 2014).

6.2 FUTURE DIRECTIONS

6.2.1 *Defining gene-sets*

The polygenic architecture of complex traits implies that a large number of genetic variants with small effect sizes influence the variation within a trait or disorder. GWAS analysis has shown repeatedly that these genetic variants are not randomly distributed over the genome. Instead, they tend to cluster in genes that share a biological pathway or cellular function (*Visscher*, 2008; *Hirschhorn*, 2009; *Lango Allen et al.*, 2010; *Ripke et al.*, 2014). This observation led to successful pathway/gene-set analyses on GWAS data as the identification of these biological pathways and molecular processes aid to increased biological insight on the disease etiology. These results also indicate the high need for accurate and relevant gene-set analyses; poorly defined gene-sets will provide misleading information on involved biological mechanisms.

Most published gene-set analyses to date have used publicly available gene-sets (derived from databases as KEGG, GO or MSigDB) (as I showed in chapter 2 & 4 (*Holmans et al.*, 2009; *O'Dushlaine et al.*, 2011; *Lips et al.*, 2015), whereas few others use in-house manually curated gene-sets (as I showed in chapter 2 & 3) (*Ruano et al.*, 2010; *Lips et al.*, 2012). The canonical pathways that can be derived from online resources are unlikely to be complete, free from error or accurate, and this may also apply to manually curated pathways, although likely to a lesser extent (*Sullivan et al.*, 2012b; *Sullivan & Posthuma*, 2015).

There is thus a high need for well-annotated gene-sets that may go beyond classic annotation of pathways as they are constructed by and KEGG or GO. However, constructing relevant gene-sets is a daunting task. In general, there are two ways in which a (gene-)set can be constructed: manual or expert knowledge based approach or a *in silico* and data-driven approach. Although expert knowledge based curation is labor intensive and does not guarantee perfection, *in silico* strategies do also not guarantee perfection since they are based on data that can

include both false-positives as false-negatives. Both will have their own constraints that should be considered. However, *in silico* strategies enable us to incorporate other data sources in association studies. Ideally, these different data sources come from the same set of individuals, but these are currently not available on a large scale. Since available techniques are constantly improved and new techniques are emerging, the future will probably enable us to systematically screen for cell-type specific events on the level of gene-expression and protein-protein interactions.

The construction of these gene-sets is not limited to pathways or even sets of biologically related genes. They can also contain a set of (non-coding) SNPs that are located in other functional elements of the genome (e.g. regulatory elements such as DHSs). Key factor in the construction is that the SNPs in the sets are located within genes or regulatory elements that are biologically related. Furthermore, the set is constructed around or at least fits the hypothesis and thus contains a high quality of relevant information. Without question, big initiatives such as ENCODE, the Epigenomic Roadmap Project, and the FANTOM project, which have the aim to comprehensively map the regulatory elements for a high number of cell-types, will provide more insights on the biology that underlies complex diseases as well as in identifying causal variants.

6.2.2 *Follow-up studies*

The identification of genes and molecular mechanisms via genetic mapping is one of the first steps that have to be taken in order to develop targeted therapies. As a next step, the genes and/or molecular mechanisms that are identified by GWAS need to be functionally validated via *in vitro* experiments. This can for example be achieved via (laborious) experiments on induced pluripotent stem cells (iPSCs), in which fibroblast cells from patients are reprogrammed to cells of interest from which cellular functions of interest can be investigated in great detail (*Lowry et al.*, 2008). Another promising technique that recently became available is CRISPR-Cas9 (*Ran et al.*, 2013). This technique has the ability to edit the DNA in a very precise and efficient manner, and can thereby provide a way to study the biological effect of one or more multiple artificial mutations *in situ* or *in vivo* in animal models. For example, Claussnitzer et al. used CRISPR-cas9 to study the effect of a predicted causal non-coding variant in the FTO gene in human primary adipose tissue. They could demonstrate that thermogenesis was turned off when the protective variant was altered into the risk variant in non-risk carriers. While

thermogenesis was rescued in individuals in which the risk variant was altered into the protective variant (*Claussnitzer et al.*, 2015).

6.2.3 *Analyses of rare variants that contribute to complex traits*

Challenges have to be tackled when extending the identification of genetic variation that contributes to complex traits to the identification of rare variants via the analysis of whole-genome sequencing (WGS) data. However, the amount of raw data that is generated through WGS for a single individual will exceed the amount of data that is seen in a typical GWAS by far. Although storage of the raw data can be questioned: when one is only interested in the genetic variation that is observed in the genome, the raw data can be discarded after filtering for an individual's genetic variation (*Stephens et al.*, 2015).

The identification of individual rare variants from sequencing data requires an enormous sample size and will probably be underpowered because even more stringent corrections for multiple testing are needed due to a much higher number of variants that are tested for association. But in a similar way as seen in gene-set analysis on GWAS data, the application of a reductionist approach, in which the genetic variation located in a single gene, or within a set of biologically related genes or other functional related DNA elements (such as DHSs and other regulatory DNA elements) is aggregated, can solve this issue of being underpowered in a similar way as shown in this thesis. In any case, to cope with the tremendous amounts of data that is generated now and will be generated in the near future, there is a high need for an infrastructure that can store and handle the computation of these data.

6.2.4 *Big Data*

Large-scale datasets are becoming the norm in genomics. A series of new technological advances, such as gene-expression analyses, genome-wide association analyses, genome sequencing, and mass spectrometry, provide ways to investigate (changes in) molecular events on a system-level. These large-scale analyses generate vast amounts of data and which has to be stored, where analyses of these data can demand huge amounts of computation time. In recent years, we have also witnessed an increase in participation between research groups because the development of the field demands participation. For instance, the studies described in Chapter 2-4 involve gene-set analysis of multiple GWAS datasets, which are only

up to 10,000 individuals in total, are collected by different research groups since no single research group is likely to collect so many data by itself. And although these dataset have a size only in the order of GB's, the computationally intensive analysis requires parallel computing. In the same way, current GWASs that include 10,000-500,000s individuals are also the result of consortia of multiple research groups, and analysis of these amounts of data require much more computation time. When current trends are expanding in the near future, it can be expected that 1) sharing data and participating in large consort and initiatives will be a necessity, 2) there will be a high need in standardization in relation to nomenclature of biological components and protocols for collecting and storing data, 3) analyses will include multiple sources of system-wide data in order to get more detailed insights in the biology of common traits, and 3) there is a high demand a large number of bioinformaticians. Therefore, a global standard on how the data should be acquired and described together with a global standard on nomenclature will aid collaboration, data integration and discovery in a future in which petabytes of data will be the norm.

Concluding, in this thesis I have showed the value of incorporation biological knowledge in a genome-wide association analysis. I expect that the integration of biological knowledge will play a major role in dissecting the genetics of complex traits, since the annotation of the human genome is currently a fast evolving field and increasingly catalogued systematically.